BRL R 1513

# BRL

AD

AD 718032

## REPORT NO. 1513

# INFORMATION AND ITS RETRIEVAL

by

Č. Masaitis

November 1970



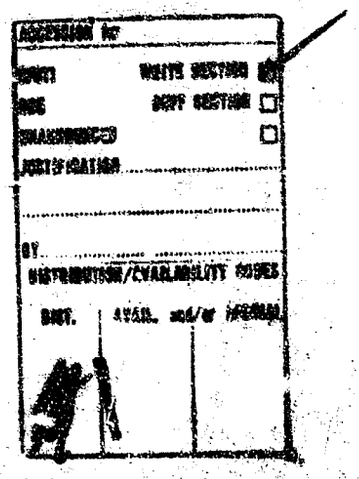




U.S. ARMY ABERDEEN RESEARCH AND DEVELOPMENT CENTER
BALLISTIC RESEARCH LABORATORIES
ABERDEEN PROVING GROUND, MARYLAND

59

Destroy this report when it is no longer needed.
Do not return it to the originator.



The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

B A L L I S T I C  R E S E A R C H  L A B O R A T O R I E S

REPORT NO. 1513

NOVEMBER 1970

# INFORMATION AND ITS RETRIEVAL

Č. Masaitis

Applied Mathematics Division



RDT&E Project No. 1T061102A14B

A B E R D E E N  P R O V I N G  G R O U N D,  M A R Y L A N D

BALLISTIC RESEARCH LABORATORIES

REPORT NO. 1513

ČMasaitis/bj
Aberdeen Proving Ground, Md.
November 1 70

# INFORMATION AND ITS RETRIEVAL

## ABSTRACT

Information is here defined abstractly as a set of highly structured elements. This structure induces a two-way classification of the elements and also other relations among them. The definition includes rules for expanding this information by incorporating new elements and for discarding elements that have become obsolete or are no longer needed. The definition of information is geared to its purpose, viz. to provide easy retrieval of known facts that are of interest to a specialist in the field.

A pilot example of such a retrieval system is included, and principles for the physical realization of the system are presented.

It is stated that a clear distinction must be made between a collection of documents and the information they contain. Likewise, the difference between recovering relevant documents and retrieving desired information is emphasized.

TABLE OF CONTENTS

# I. INTRODUCTION

This report discusses problems in the field of information storage and retrieval. For years numerous libraries, organizations and individuals have devoted their resources and intelligence to the practice and methodology of these problems. Many methods have been proposed and discussed; some of them have been tried on a smaller or larger scale.

Nevertheless we feel that this report contains some aspects that are not a mere repetition or reformulation of ideas discussed elsewhere. First of all, we make a clear distinction between documents and information, and we consider the problem of storage and retrieval of the latter. Secondly, we propose that information systems should be constructed by professionals. This is just the opposite of an attempt to mechanize the classification of documents. Thirdly, we do not consider available technical devices that can be used to record and recover information. Rather we discuss the functions that should be performed by a retrieval system.

Certainly, documents such as books, reports, charts, and maps are a source of information. Therefore it is extremely important to store, to classify, and to catalog them. However we do not suggest any novel approach to this problem, nor do we consider any methods to improve existing means for storing and retrieving documents. Our concern is with information, and we contend that documents do not constitute information. In fact, quite a few documents do not contain any information, in spite of their eye-catching titles and summaries. Hence no sorting, classifying, and labeling of such documents will create an information system. For this reason, besides the others, Baxendale's [1]* counting of "non-general words" and Borko's and Bernick's [2] linear regression formula can yield only a collection of documents that is hardly relevant to one's request for information. In many cases there

*References are listed on page 54.

is, indeed, a very low correlation between the frequency of certain words in a document and the information contained or not contained in it, or between Luhn's ill-defined notions [3] with their statistics and what is really said in the document. You may find a paper that mentions electron-volt more than a hundred times and yet says nothing about its relation to engineering units of energy. Yet another document may mention electron-volt only once and right there may express it in BTU's.

We cannot separate information and intelligence. Since neither a file nor an electronic system can be intelligent, such devices cannot automatically retrieve desired information from a heap of relevant and irrelevant documents. Nor can a clerk fare any better when he encounters technical language completely foreign to him. Consequently, we need a professional designer to extract information from its source and to present it to the user by the most efficient means, most efficient over a span of time and a number of requestors.

Here the meaning of "most efficient" is not with respect to available means but with respect to the goal of information retrieval. We did not survey all the forms of sea-going ships and all the shapes of winged aircraft in a search for a conveyance to carry us to the moon. Instead, we looked first for what it takes to go there, and then we examined whether it was feasible and how much it might cost. The same approach is proposed here to the problem of information retrieval: consider first what is needed to achieve the goal, and afterwards examine the feasibility of any proposed approach. Of course, there is more than one way to build and to stage a spaceship. We do not claim that the approach to retrieval of information discussed in this report is necessarily the best or even one of the better ones. However, we feel that our plan of attack is one of the right ones.

We propose first to define information. The definition must be constructed according to the purpose of the information system. Such a definition is presented in Section II. It is followed by an example in Section III that illustrates the abstract definition in concrete terms.

Section IV attempts to present the questions that summarize various requests for information on a subject. Furthermore, Section IV discusses ways to obtain answers to these questions. We hope this discussion provides motivation for the definition of information presented in Section II and also establishes a link with Section V which presents principles for the physical realization of an information system. Section VI attempts to compare the effort needed to construct the proposed system with the need for efficient retrieval of information.

## II. INFORMATION

Frequently a man busy with his research task would like to obtain needed facts without having to search for hours through various references. However, at other times he may want to range through the field of his specialty without looking for particular information. He may just want to wander around and see what is going on in his field of research. Hence, even if specialized and efficient information systems were developed, we still would need libraries. It seems that two types of libraries would be desirable. First we should have conventional libraries serving users in their neighborhood. Here you could get a hard copy of a document to browse through in the familiar manner that you learned long ago, beginning with kiddies' coloring books. The other type should be a library of the future, one that does not exist now and, most likely, is not feasible yet.

In this library of the future books, magazines, journals, maps, reports, charts, graphs, and any other conceivable types of documents should be put on some kind of ultramicrofiche, all of which would be mounted on a superprocessor with thousands of parallel channels that would provide a simultaneous and immediate read-out of many thousands of documents. The channels should be connected to remote access stations at which documents of your choice would be displayed on screens and also copied if necessary.

This central library should contain a considerable portion of the documents in all fields of knowledge that are published or accepted for publication. Of course, not all documents should be included. Let us recall that the Library of Congress, holding probably the world's largest collection of documents, selects only about two million books and other items out of an annual supply of a few billion pieces that reach the Library. The library of the future should be no less selective. It should not be a dump for every piece of printed matter.

Immediate access to the material in this library, perhaps, would allow us to do away with many technical journals. Editors of these journals could become a part of the library team that selects the documents to be stored. Then a researcher in any field would not have to wait one to three years to read what new thing has been found by his fellow researcher who works in the next-door institution, as he has to wait now with the present methods of refereeing, editing, publishing, and distributing of technical material.

Present methods for classifying and indexing documents would hardly be adequate for such a superlibrary of the future. Therefore an adequate number of remote access stations should be rigged to record the steps taken by users in an attempt to retrieve documents of a desired type. One should record both the ingenuity and the frustration of individual users. These data should be collected by field of knowledge, type of documents, and by class of users. Careful and continuous analysis of such data would allow a continuous improvement of the software and the hardware of the system. It also would serve as guidance in the future selection of documents for storage in the library. Obviously, no system can add new material forever without removing something that is already in the system. Collected data that would show what is being used and how would also provide criteria for removal of material.

Similar data collected for any existing document retrieval system would reveal its weak and strong points and at the same time would help find ways for improvement. An effort spent in collecting such data

may be very worthwhile. There is no way to study reality other than to conduct actual observations of facts. This includes the operation of a document retrieval system, as well.

We are so accustomed to books, journals, and conventional libraries that it seems absolutely impossible to do away with them. This report is primarily concerned with information, not documents. Nevertheless the preceding paragraphs about libraries were included simply to express our prejudice against spoon-fed information. Seemingly, many researchers still feel - at least subconsciously - that they should read for themselves everything published in their field just as researchers felt fifty or a hundred years ago. Or at least they feel that they should be able to read anything they choose, including trivial verbiage published in their field of knowledge. However, this may not be the most favorable condition for the advancement of research nor the most advantageous situation for an individual researcher. It is quite possible that sufficiently diversified and adequately complete information systems containing only significant results - significant by the judgment of selected groups of scientists - would satisfy the needs of a research community, and hence there would be no need for a superlibrary as described above. Only because of the prejudice of the author for libraries and because of his do-it-yourself attitude was the above digression made.

This digression may also serve to emphasize the contrast between an all-inclusive (or almost all-) library and a specialized information system that should provide up-to-date knowledge accessible to a specialist only and that could help him to plan further steps into the unknown and also would reduce duplications. In other words, an information system should be highly specialized, and it should be able to respond to the queries put forth by a specialist in the field. Indeed, there is no point in building an information system on, say, partially ordered spaces that would provide adequate answers to questions asked by a lawyer who never learned his high school algebra. Only a specialized information system designed for a specialist can be useful in research.

Examples of specialized systems that are concerned with information and not with documents are the original Chemical Information and Data System [4] and American Airlines' SABRE Electronic Reservation System [5]. Differences in the structure and in the methods of implementation of these two systems indicate that an information system depends on the type of information involved and on the purpose of retrieval. Most likely, no single description of an information system could cover all possible areas. Each field of knowledge, or at least each family of related branches of science, may need a different definition of information and also different methods for storage and retrieval. Development of several information systems may provide material for study of common principles and may lead to discovery of more efficient methods of retrieval.

Branches of mathematics are the best organized logical systems. Therefore it seems that information in mathematical disciplines may be easier to study and a desired mode of retrieval may be simpler to describe than in other fields of knowledge such as, for instance, political science or philosophy, where an appeal to the emotion of the reader by the rhetorical garb of presentation frequently outweighs the logic and the reasoning of the argument. Consequently, we choose the field of mathematics for our study of the principles of information.

A mathematical discipline consists of undefined terms, among many other things. Some of these terms are primitive concepts such as lines or points in Euclidean geometry. Other concepts are primitive relations. For instance, "every pair of distinct points defines a single line" is an undefined relation between a pair of points and a line in a Euclidean space.

We start with the set of all such undefined terms in a given system. This set need not and should not be as small as logically possible, i.e. it should contain terms that cannot be defined as well as terms that are commonly known to a specialist in the field and hence need not be defined. A person or a group of persons that would

undertake the development of an information system should decide which terms are "commonly known" and which are not. For short we will call such a person or a group a designer.

The designer should also choose what to include in the system and what not to. New developments in the field should be screened by him and either included in the information system or rejected as trivial, irrelevant, or not new. This is an arbitrary procedure. It is possible that some gems of discovery will never appear in the system and some trivia will be included. Well, nothing is perfect. Nor should one strive to construct a perfect information system. It seems that the most desirable information retrieval system would be an intelligent expert who possesses the experience of several scholars and who can read, screen, and evaluate as much new material as a score of diligent students combined. However, such a superexpert still would have his prejudices, likes, and dislikes, his oversights, and exaggerations. Yet we would be happy to have such a consultant, especially if the only credit required for his service would be what a researcher now gives to the library.

We imagine our information system as a sort of a superhandbook that is up-to-date, complete, and so well organized that any information contained in it can be retrieved with great ease. At present we have competing handbooks and monographs in many fields of knowledge. There is no reason why competing information systems should not be constructed. Similar competing commercial information systems are visualized in the future. Some of them are described in [6]. Competition would encourage us to design information systems that suit the needs of users instead of merely the taste of the designer.

After these remarks we return to our definition of an information system. As we choose the undefined terms to be included in our information, we should also consider synonyms. Thus, we introduce elements that consist of two components: a main term and the set of its synonyms. Furthermore we introduce a label for each element. We

call this label the "address of the element". It consists of three components: (d,o,i). The symbol i is defined below. We abbreviate this to $d_i^o$. This symbol will be used with two different meanings: the address of an element and the element itself, i.e. we write $d_i^o = (\tilde{d}_i^o, (d_i^o))$, where $\tilde{d}_i^o$ is the main term and $(d_i^o)$ is the set of synonyms. Thus, an element $d_i^o$ of our system that corresponds to an undefined notion or an undefined relation has two components: the first component is the main term, and the second component is the set of synonyms, which may be empty. We call the elements of $(d_i^o)$ primary synonyms. The collection of all such pairs (together with their addresses) is denoted by ${}_oD$. This collection we call information of level zero. We will also use the symbol $I_o = {}_oD$.

The first two components of an address are d and o for every element of ${}_oD$. The third component, i, is a real number of the form $\frac{a}{2^n}$, where a is a positive integer and $a < 2^n$. We form an alphabetic list of all the main terms and assign a third component to each element in such a way that these components form a monotone strictly increasing sequence. The alphabetic list of main terms together with their addresses is called the ${}_oD$-dictionary.

Besides the set $(d_i^o)$ of primary synonyms we also consider the set $(\tilde{d}_i^o)$ of secondary synonyms. This set may be empty or not. It contains some rare or old synonyms such as, for instance, "analysis situs" for topology and some likely misspellings of the main term or primary synonyms. We form an alphabetic list that includes all main terms, primary and secondary synonyms, each labeled with the respective address. We call this list the ${}^oD$-dictionary (or a $D_o$-dictionary). As stated above the set ${}_oD$ (or $I_o$) is an information of level zero. The set $I_o$ together with ${}_oD$ - and ${}^oD$-dictionaries, with some means of recording the elements and with the rules for adding new elements, for deleting some of them, and for retrieving constitutes an information system of level zero. The operation rule that provides the address of an element when the main term or any of its synonyms is specified is

called the element translator. This rule may be simply a look-up in the ${}^{o}D$-dictionary.

We define information of level n inductively. Suppose that information $I_{n-1}$ of level n-1 has been defined. We assume that $I_{n-1} = D_{n-1}UP_{n-1}UC_{n-1}US_{n-1}UT_{n-1}$, the logical sum of five sets to be described below. For n = 1 we set $D_{n-1} = D_o = {}_oD$, $P_o = C_o = S_o = T_o = \phi$ (the empty set). Let $[d_i^n]$ be a verbal statement of a definition of a concept $\tilde{d}_i^n$ expressed in terms of words contained in the $D_{n-1}$-dictionary or in terms of their semantic equivalents (such as plural or possessive form, past tense, etc.) together with logical connectives, such as "or", and quantifiers such as "there exist", and also with any non-technical terms (non-technical with respect to the particular field of information) such as "of", "for", etc. In other words $[d_j^n]$ is a definition of the term $\tilde{d}_i^n$ freely chosen by the designer of the information system with the single restriction that this definition must not use any technical terms (no matter how common and simple) except those contained in the $D_{n-1}$-dictionary.

We assume that every definition consists of a generic term t and a qualifying statement. The term t and any technical term used in a qualifying statement are in the $D_{n-1}$-dictionary. Hence each such term is associated with a corresponding address $d_j^k$, $k < n$. We denote by ${}^nD_i$ the set whose single element is the address of t and by $D_i^n$ the set of addresses of all the terms contained in the qualifying statement. We also choose the set $(d_i^n)$ of primary synonyms of the term $\tilde{d}_i^n$ and the set $(\tilde{d}_i^n)$ of its secondary synonyms. Next we choose a document reference which, in our judgment, contains satisfactory discussion of all or some of the following aspects: motivation of the definition and of the choice of the term, a proof that the definition is non-vacuous and meaningful, and a list of other pertinent references. We denote this reference by $\{d_i^n\}$, i.e. $\{d_i^n\}$ consists of an author's name, the title of his work, page, date, etc. Finally, we assign an integer $\bar{d}_i^n$ to the term $\tilde{d}_i^n$ and

call that integer an index of rescission. This is a control parameter for removing elements by the rules described below. Thus, we construct a septemplet $(\tilde{d}_i^n, (d_i^n), {}^nD_i, D_i^n, [d_i^n], \{d_i^n\}, \bar{d}_i^n)$. We assign an address to this element as follows: the first component is d, the second component is obtained by adding one to the maximum of the second component of all addresses contained in ${}^nD_i$ and $D_i^n$; the third component i is assigned in a similar fashion as for the elements of ${}_oD$. The collection of all these septemplets, together with their addresses, where the second component of the address is a fixed value n, is the set ${}_nD$. The alphabetical list of all the terms $\tilde{d}_i^n$ together with their addresses is the ${}_nD$-dictionary. We also have the ${}^nD$-dictionary that includes the main terms $\tilde{d}_i^n$ and all their synonyms along with the address of each. Of course, we assume again that a translator, i.e. a rule for obtaining the address of any main term or its synonym, is available.

Besides definitions, a discipline in mathematics contains postulates, conjectures, and theorems. The logical form of a theorem is an implication, i.e. the main logical connective in a theorem is "if" or "if and only if". We call the first type of theorem unsymmetric or of the t-type, and the second type of theorem symmetric or of the s-type. Obviously, existence theorems such as "there exists a unique solution of $Lu = 0$" can readily be expressed in a form of implication such as: "If L is an operator with the property P, then there exists a unique element u such that $Lu = 0$". Similarly, postulates and conjectures can be expressed in a form of implication. Thus, for instance, the Peano's axiom: "One is a natural number" can be stated as: "If an element is one, then it belongs to the set of natural numbers".

We define the elements of information of level n that correspond to symmetric theorems in a fashion very similar to that used in our definition of elements in the set ${}_nD$. Namely, we denote by $[s_i^n]$ a verbal statement of a symmetric theorem. The symbol $\tilde{s}_i^n$ is an identification symbol such as "Cauchy Theorem" etc. If the theorem has no name, the symbol $\tilde{s}_i^n$ is simply its address in the information. The

sets $(s_i^n)$ and $(\tilde{s}_i^n)$ are the sets of primary and secondary synonyms. The main connective in a theorem divides it into two parts. We choose one of these parts as the antecedant (it is a symmetric theorem) and the other as the consequent. The addresses of the terms contained in the antecedant constitute the set ${}^nS_i$, and those belonging to the terms in the consequent belong to the set $S_i^n$. Of course, these are the addresses of terms in the $D_{n-1}$-dictionary. Again we choose a reference $\{s_i^n\}$ and a rescission index $\bar{s}_i^n$ in a fashion similar to that used in the case of elements of ${}_nD$. Thus, we obtain $s_i^n = (\tilde{s}_i^n, (s_i^n), {}^nS_i, S_i^n, [s_i^n], \{s_i^n\}, \bar{s}_i^n)$. The collection of all such elements in the information is denoted by ${}_nS$.

Similarly, we define the set ${}_nT$ of unsymmetric theorems, the set ${}_nP$ of postulates, and the set ${}_nC$ of conjectures. Elements of these sets consist of seven components each and have addresses of the type $t_i^n$, $p_i^n$, and $c_i^n$, respectively. We also construct an ${}_nS$-dictionary and an ${}^nS$-dictionary, similar to the ${}_oD$- and the ${}^oD$-dictionaries. Similarly, we have dictionaries for ${}_nT$, ${}_nP$, and ${}_nC$. Now the information of level n is $I_n = I_{n-1} \cup {}_nD \cup {}_nS \cup {}_nT \cup {}_nP \cup {}_nC$. We also define the sets $D_n = \bigcup_{k=o}^{n} {}_kD$, $S_n = \bigcup_{k=o}^{n} {}_kS$, $T_n = \bigcup_{k=o}^{n} T_k$, $P_n = \bigcup_{k=o}^{n} P_k$, and $C_n = \bigcup_{k=o}^{n} {}_kC$.

Frequently we will discuss the elements of information without specifying their type. Hence we introduce the notation $a_i^k$, $(a_i^k)$, ${}^kA_i$, ${}_kA$, $A_k$, etc with a denoting either d(definition), or s(symmetric theorem), or t, or, p, or c. Similarly, A stands for any of the five capital letters. D, S, T, P, C. Similarly, $b_j^\ell$ etc. denote generic elements of the information. We also assume that we have $A_n$-dictionaries that are alphabetical lists of main terms and all the synonyms contained in the elements of $A_n$ together with their addresses. We also construct an $I_N$-dictionary that is obtained by combining $A_N$-dictionaries for all five values of A. Here N denotes the highest level of the elements

contained in the information. We call this $I_N$-dictionary the main dictionary of the information.

The structure of the elements of information induces certain relations. We say that $a_i^k$ and $b_j^\ell$ are in a relation of order zero to each other if there exists an element $a_z^m$ such that $a_i^k, b_j^\ell \in {}^mA_z \cup A_z^m$. The element $a_z^m$ is in a left relation of order one to the element $a_i^k$, provided $a_i^k \in {}^mA_z$; $a_z^m$ is in a relation of order one to the element $a_i^k$, provided $a_i^k \in {}^mA_z \cup A_z^m$. If $a_z^m$ is in a (left) relation of order one to $a_i^k$, then we say that $a_i^k$ is in a (left) relation of order minus one to $a_z^m$. The verbal statement $[a_z^m]$ is called a natural relation between $a_i^k$ and $a_z^m$ (of order one or order minus one), and also it is called a natural relation of order zero between $a_i^k$ and $b_j^\ell$.

Besides these natural relations of order minus one, zero, or one, we define relations of higher (positive or negative) order inductively as follows. Suppose that $a_i^k$ is in a relation of order $n_1$ to the element $a_z^m$ and that $a_z^m$ is in a relation of order $n_2$ to the element $b_j^\ell$. Then we say that $a_i^k$ is in a relation of order $n_1 + n_2$ to $b_j^\ell$ provided $n_1 n_2 > 0$.

A finite sequence of elements $\ell_0, \ell_1, \ldots, \ell_k$ is called a k-chain of elements provided each element of the sequence (except the last one) is in a relation of order one to the next element. Similarly, the sequence $\ell_0, \ell_1, \ldots, \ell_k$ is called a minus k-chain if every element of the sequence (except the last one) is in a relation of order minus one to its neighbor on the right. The sequence of natural relations of order one (or minus one, respectively) between the successive elements of a k-chain (minus k-chain) of elements is called a k-chain (minus k-chain) of relations. It is easy to prove the following theore..

Suppose $\ell_o$ is in a relation of order k(minus k) to $\ell_k$. Then there exists a k-chain (minus k-chain) of elements that starts with $\ell_o$ and ends with $\ell_k$.

We say that this is a k-chain (or minus k-chain) of elements from $\ell_o$ to $\ell_k$. Similarly, the corresponding k-chain of relations is a k-chain of relations from $\ell_o$ to $\ell_k$.

In Sections IV and V an information retrieval process is expressed as a display of components of elements and k-chains of elements or relations. The same process could be expressed in terms of operations on certain matrices. One can construct a matrix M that represents the structure of the information system (including the relations). For every n=0,1,2,...,N we put ${}_nA$-dictionaries one after the other in the sequence ${}_nD$, ${}_nP$, ${}_nS$, ${}_nT$, and ${}_nC$ to form an n-list of elements. Then we put the n-lists in a sequence: n=0,1,2,...,N and number the elements successively from 1 to card $I_N$. Now we define the elements of a matrix M as follows: $m_{ij} = 1$ if $j \geq i$ and the i-th element of $I_N$ is in a relation of order minus one to the j-th element. Otherwise $m_{ij} = 0$ for $j \geq i$. The elements $m_{ij}$ for $j < i$ are defined by the relation $m_{ij} = - m_{ji}$, i.e. M is skew-symmetric.

This definition of M can be used to compare our information system with the analysis presented in [7]. Since M is skew-symmetric it is sufficient to consider only its upper-triangular part. This part can be partitioned into rectangular submatrices that represent structural relations of $J_k = I_k \setminus I_{k-1}$ (the set theoretical difference of $I_k$ and $I_{k-1}$) with $D_{k-1}$. Since the elements of $J_k$ are related only to the elements of the form $d_j^m$, one can omit the rows of every matrix obtained by partitioning that correspond to the elements of $I_{k-1}$ other than $d_j^m$. We denote the resulting matrix by $M_k$. Analysis of these matrices may be employed to determine connectedness of the information, maximum length of k-chains between two given elements, and many other questions about the structure of the information. These problems are interesting from an abstract view point. However, when passing from $I_N$ to the matrix

M we lose the most important aspects of $I_N$, namely the components $[a_i^k]$ of the elements in $I_N$. Without these verbal statements the system loses its practical value as an information retrieval system.

In summary, an information system consists of elements that are septemplets whose successive components are: name(identification), set of primary synonyms, set of elements in antecedant (or generic term), set of elements in consequent (or qualifying statement), verbal description of the element, bibliographical reference, and rescission index. These septemplets are provided with labels or addresses and are grouped into sets that constitute information of various levels, beginning with level zero up to level N. The highest level N may increase or decrease with the life of the system. This is discussed in Section V. The elements of information of level zero are pairs instead of septemplets. Furthermore, relations of order $k(k = -N; -N+1,\ldots,0,1,2,\ldots,N)$ among the elements of $I_N$ and k-chains of elements and of natural relations are defined. Furthermore, we construct various dictionaries.

This definition of information was chosen with a discipline in mathematics in mind. It is certainly not universally applicable. However, this definition may be suitable for many other technical and scientific disciplines. After all, one can say the same about many research fields as is said on p. 186 of [8]: "Economic theory, not unlike other theories, consists of three basic elements - definitions, assumptions, and conclusions". Thus, in many fields observed facts, data, and commonly used technical terms can be assigned to $I_0$, i.e. can be treated as information of level zero. Definitions can be included in the sets ${}_kD$ in the manner described above. Empirical relations established between observed facts and collected data constitute postulates of various levels. Hypotheses are conjectures. Theory and its deductive conclusions are equivalent to theorems in mathematics. Queries for information also can be interpreted in terms of the types of questions in mathematics as presented in Section IV below.

## III. EXAMPLE

We illustrate the definition developed in the preceding section by an example that has been constructed from the discussion contained in Section I, Chapter 1, in [9]. The set ${}_{o}D$ of this example is rather small. It contains only 29 elements, and these are listed in Table I. This Table is our ${}_{o}D$-dictionary. Obviously, in a specialized information

Table I

$d_1^o$ = addition,
$d_2^o$ = cardinal number,
$d_3^o$ = cartesian product,
$d_4^o$ = collection,
$d_5^o$ = countable,
$d_6^o$ = difference of sets,
$d_7^o$ = element of,
$d_8^o$ = empty set,
$d_9^o$ = equal,
$d_{10}^o$ = finite,
$d_{11}^o$ = first element of the ordered pair,
$d_{12}^o$ = identity mapping of a set A onto set A,
$d_{13}^o$ = indexing set,
$d_{14}^o$ = integer,
$d_{15}^o$ = intersection,
$d_{16}^o$ = minimum,
$d_{17}^o$ = not an element of,
$d_{18}^o$ = not equal,
$d_{19}^o$ = one-one,
$d_{20}^o$ = ordered pair,
$d_{21}^o$ = parentheses (),
$d_{22}^o$ = proper subset,
$d_{23}^o$ = second element of the ordered pair,
$d_{24}^o$ = sequence,

Table I (Continued)

$d^0_{25}$ = set,
$d^0_{26}$ = set of all,
$d^0_{27}$ = set of positive integers,
$d^0_{28}$ = subset,
$d^0_{29}$ = union.

system terms like single-valued mapping, range, and binary operation need not be defined, i.e. they should be included in ${}_0D$. Instead, we have chosen to consider these terms and many other commonly known concepts as elements of informations of higher level since this example is intended only to illustrate the principles outlined earlier.

With this choice of ${}_0D = I_0$ the total information contained in Section I of [9] extends up to $I_{12}$. The sets $P_{12}$ and $C_{12}$ are empty (there are no postulates or conjectures in this section). The sets $S_8$ and $T_8$ are also empty. Some sets such as ${}_1D$, ${}_3D$, ${}_8D$, and others contain only one element each. The next largest set after ${}_0D$ is ${}_5D$ with twelve elements. Table II can be interpreted as a collection of ${}_kD$-dictionaries, k=1,2,...,11. This Table contains the main terms of each set ${}_kD$ in alphabetical order. The left column contains the addresses of the corresponding elements. For simplicity the subscripts i in these addresses are written with a common denominator (power of 2) that is omitted. Thus, $d^1_1$ should really be $d^1_{\frac{1}{2}}$, and $d^2_3$ should be $d^2_{\frac{3}{2}}$. Similar simplification is adopted in Table I and also in other notation of this example. Table III contains the sets $(d^k_i)$ of primary synonyms that are not empty. In this example all the sets of secondary synonyms are empty.

Table II

| | |
|---|---|
| $d_1^1$ | Single valued mapping |
| $d_1^2$ | Binary operation |
| $d_2^2$ | Inverse mapping |
| $d_3^2$ | Range of a single valued mapping |
| $d_1^3$ | Halfgroupoid |
| $d_1^4$ | Compatible collection of halfgroupoids |
| $d_2^4$ | Disjoint collection of halfgroupoids |
| $d_3^4$ | Divisor in a halfgroupoid |
| $d_4^4$ | Groupoid |
| $d_5^4$ | Homomorphism of a halfgroupoid into a halfgroupoid |
| $d_6^4$ | Order of a halfgroupoid |
| $d_7^4$ | Subhalfgroupoid |
| $d_1^5$ | Countable halfgroupoid |
| $d_2^5$ | Divisor chain in a halfgroupoid |
| $d_3^5$ | Endomorphism of a halfgroupoid |
| $d_4^5$ | Extension of a halfgroupoid |
| $d_5^5$ | Finite halfgroupoid |
| $d_6^5$ | Homomorphism onto |
| $d_7^5$ | Intersecting collection of halfgroupoids |
| $d_8^5$ | Isomorphism |
| $d_9^5$ | Prime element |
| $d_{10}^5$ | Subgroupoid |
| $d_{11}^5$ | Subhalfgroupoid closed in the halfgroupoid |
| $d_{12}^5$ | Union of a compatible collection of halfgroupoids |
| $d_1^6$ | Automorphism |
| $d_2^6$ | Complete extension of a halfgroupoid |
| $d_3^6$ | Disjoint collection of groupoids with amalgamated subgroupoids |
| $d_4^6$ | Extension chain of halfgroupoids |

Table II (Continued)

| | |
|---|---|
| $d_6^6$ | Finite divisor chain over a halfgroupoid |
| $d_8^6$ | Imbeddable halfgroupoid |
| $d_7^6$ | Induced homomorphism |
| $d_8^6$ | Intersection of an intersecting collection of halfgroupoids |
| $d_9^6$ | Open extension of a halfgroupoid |
| $d_1^7$ | Complete extension chain of halfgroupoids |
| $d_2^7$ | Halfgroupoid free over its subhalfgroupoid |
| $d_3^7$ | Length of a finite divisor chain |
| $d_4^7$ | Maximal extension chain |
| $d_5^7$ | Open extension chain |
| $d_1^8$ | Subhalfgroupoid generates a subhalfgroupoid |
| $d_1^9$ | Freely generated halfgroupoid |
| $d_1^{10}$ | Free basis of a halfgroupoid |
| $d_2^{10}$ | Free product of a disjoint collection of halfgroupoids |
| $d_3^{10}$ | Generalized free product of a compatible collection of halfgroupoids |
| $d_1^{11}$ | Free halfgroupoid |
| $d_2^{11}$ | Generalized free product of a disjoint collection of groupoids with amalgamated subgroupoids |

Table III

$(d_8^6)$ = (Imbedded halfgroupoid, Imbedding of a halfgroupoid),

$(d_7^6)$ = (Homomorphism extends a homomorphism, Extension of homomorphism),

$(d_1^8)$ = (Subhalfgroupoid of halfgroupoid generated by a subhalfgroupoid),

$(d_1^9)$ = (Halfgroupoid freely generates a halfgroupoid).

Table IV lists the verbal statements $[d_i^k]$.

Table IV

$[d_1^1]$ = Single valued mapping f of a set A into a set B (f,A,B) is a subset of cartesian product A × B such that (a,b), (a,c) $\epsilon$ f implies b = c.

$[d_1^2]$ = Binary operation f on a set G is a single valued mapping (f, G × G, G).

$[d_2^2]$ = Inverse mapping of a single valued mapping (f,A,B) is a single valued mapping $(f^{-1},B,A)$ such that (b,a) $\epsilon$ $f^{-1}$ if and only if (a,b) $\epsilon$ f.

$[d_3^2]$ = Range of a single valued mapping (f,A,B) is a subset R(f) of A such that a $\epsilon$ R(f) implies that there exist b $\epsilon$ B and (a,b) $\epsilon$ f.

$[d_1^3]$ = Halfgroupoid $\mathcal{J}$ is an ordered pair (J,f) whose first element is a set J and the second element is a binary operation f on the set J.

$[d_1^4]$ = Compatible collection of halfgroupoids $\{(H_\alpha,f_\alpha)|\alpha \epsilon A\}$ is a set of halfgroupoids $(H_\alpha,f_\alpha)$ with $\alpha \epsilon A$ where A is an indexing set such that $\alpha,\beta \epsilon A$ and $(a,b) \epsilon R(f_\alpha) \cap R(f_\beta)$ imply that $f_\alpha(a,b) = f_\beta(a,b)$.

$[d_2^4]$ = Disjoint collection of halfgroupoids $\{(H_\alpha,f_\alpha)|\alpha \epsilon A\}$ is a set of halfgroupoids $(H_\alpha,f_\alpha)$ with $\alpha \epsilon A$ where A is an indexing set such that $\alpha,\beta \epsilon A$ and $H_\alpha \cap H_\beta \neq \phi$ implies that $\alpha = \beta$.

$[d_3^4]$ = If a, b, c $\epsilon$ H then a and b are divisors of c in a halfgroupoid (H,f) provided c = f(a,b).

$[d_4^4]$ = Groupoid $\mathcal{J}$ is a halfgroupoid (J,f) such that R(f) = J × J.

$[d_5^4]$ = Homomorphism of $\mathcal{K}$ = (H,f) into $\mathcal{J}$ = (G,g) is a single valued mapping $(\theta,H,G)$ such that a, b, c $\epsilon$ H and c = f(a,b) imply $\theta(c) = g(\theta(a),\theta(b))$.

$[d_6^4]$ = Order of a halfgroupoid (H,f) is the cardinal number of H.

$[d_7^4]$ = Let $\mathcal{K}$ = (H,h) be a halfgroupoid and let J ⊂ H. Let g = $\{(a,b)|a,b \epsilon J$ and $(a,b) \epsilon h\}$. Then $\mathcal{J}$ = (J,g) is a subhalfgroupoid of $\mathcal{K}$.

Table IV (Continued)

$[d_1^5]$ = Halfgroupoid is countable when its order is countable.

$[d_2^5]$ = Let $\mathcal{K} = (H,h)$ be a halfgroupoid and $\{a_i\}$ be a sequence of elements of H. If $i \in N$ implies that $a_i$ is divisor of $a_{i+1}$ then $\{a_i\}$ is a divisor chain in $\mathcal{K}$.

$[d_3^5]$ = Endomorphism of a halfgroupoid $\mathcal{K}$ is homomorphism of $\mathcal{K}$ into $\mathcal{K}$.

$[d_4^5]$ = Extension of a halfgroupoid $\mathcal{J} = (J,g)$ is a halfgroupoid $\mathcal{K} = (H,h)$ such that $J \subset \mathcal{K}$ and $R(h) \subset J \times J$.

$[d_5^5]$ = Halfgroupoid is finite when its order is finite.

$[d_6^5]$ = Homomorphism $\theta$ of a halfgroupoid $\mathcal{J}$ onto a halfgroupoid $\mathcal{K}$ is a homomorphism of $\mathcal{J}$ into $\mathcal{K}$ such that $\theta(J) = H$.

$[d_7^5]$ = Intersecting collection of halfgroupoids $\{\mathcal{K}_\alpha | \alpha \in A\}$ is a compatible collection of halfgroupoids $\mathcal{K}_\alpha = (H_\alpha, f_\alpha)$ such that $\bigcap_{\alpha \in A} H_\alpha \neq \phi$.

$[d_8^5]$ = Isomorphism $\theta$ of halfgroupoids $\mathcal{J} = (J,g)$ and $\mathcal{K} = (H,h)$ is homomorphism of $\mathcal{J}$ into $\mathcal{K}$ that is a one-to-one single valued mapping of J onto H.

$[d_9^5]$ = Let $\mathcal{K} = (H,h)$ be a halfgroupoid and $a \in H$. Then a is prime if it has no divisors in $\mathcal{K}$.

$[d_{10}^5]$ = Subgroupoid of a halfgroupoid $\mathcal{K} = (H,h)$ is a subhalfgroupoid $\mathcal{J} = (J,g)$ such that $\mathcal{J}$ is groupoid.

$[d_{11}^5]$ = Subhalfgroupoid $\mathcal{J} = (J,g)$ of halfgroupoid $\mathcal{K} = (H,h)$ is closed in $\mathcal{K}$ if $a,b \in J$ and $(a,b) \in R(h)$ imply that $(a,b) \in R(g)$.

$[d_{12}^5]$ = Union of a compatible collection of halfgroupoids $\{(H_\alpha, h_\alpha)\}$ is a halfgroupoid $\mathcal{J} = (J,g)$ such that $J = \cup H_\alpha$ and $g(a,b) = c$ for $a, b, c \in J$ iff there exists $\alpha$ such that $h_\alpha(a,b) = c$.

$[d_1^6]$ = Automorphism of a halfgroupoid $\mathcal{K}$ is an isomorphism of $\mathcal{K}$ and $\mathcal{K}$.

$[d_2^6]$ = Complete extension of a halfgroupoid (J,g) is an extension (H,h) such that $J \times J \subset R(h)$.

Table IV (Continued)

$[d_3^6]$ = Disjoint collection of groupoids with amalgamated subgroupoids is a disjoint collection of groupoids $\{\mathcal{J}_\alpha \mid \alpha \in A\}$ with a set of groupoids $\{\mathcal{J}_{\alpha\beta} = (J_{\alpha\beta}, j_{\alpha\beta}) \mid \alpha,\beta \in A \text{ and } \mathcal{J}_{\alpha\beta} \subset \mathcal{J}_\alpha\}$ such that $\mathcal{J}_{\alpha\beta}$ and $\mathcal{J}_{\beta\alpha}$ are isomorphic and there exist a compatible collection of groupoids $\{\mathcal{K}_\alpha = (K_\alpha, k_\alpha)\}$ such that for every $\alpha \in A$ there exist an isomorphism $\varphi_\alpha$ of $\mathcal{J}_\alpha$ and $(K_\alpha, k_\alpha)$ and $\varphi(\mathcal{J}_{\alpha\beta}) = K_\alpha \cap K_\beta$.

$[d_4^6]$ = Extension chain of halfgroupoids is a sequence $\{\mathcal{H}_i\}$ of halfgroupoids such that $i \in N$ implies $\mathcal{H}_{i+1}$ is an extension of $\mathcal{H}_i$.

$[d_6^6]$ = Divisor chain $\{a_i\}$ in a halfgroupoid $\mathcal{H}$ is finite over $\mathcal{H}$ if there exists an integer k such that for every natural number $p, a_{k+p} = a_k$.

$[d_8^6]$ = Halfgroupoid $\mathcal{J}$ is imbeddable in halfgroupoid $\mathcal{H}$ if there exist a subhalfgroupoid $\mathcal{K}$ of $\mathcal{H}$ isomorphic with $\mathcal{J}$.

$[d_7^6]$ = Let $\varphi$ be a homomorphism of a halfgroupoid $\mathcal{J} = (J,g)$ into a halfgroupoid $\mathcal{H} = (H,h)$ and let $\mathcal{K} = (K,k)$ be a subhalfgroupoid of $\mathcal{J}$. Let $\theta$ be a single valued mapping from K into H such that for every $a \in K, \theta(a) = \varphi(a)$. Then $\theta$ is a homomorphism induced by $\varphi$.

$[d_8^6]$ = Intersection of intersecting collection of halfgroupoids $\{(H_\alpha, h_\alpha) \mid \alpha \in A\}$ is a halfgroupoid $\mathcal{J} = (J,g)$ such that $J = \bigcap_{\alpha \in A} H_\alpha$ and for $(a,b) \in (J \times J) \cap R(h_\alpha), g(a,b) = h_\alpha(a,b)$.

$[d_9^6]$ = Open extension of a halfgroupoid $(J,g)$ is an extension $(H,h)$ such that $(a,b) \in R(h)$ and $h(a,b) \in J$ implies that $(a,b) \in R(g)$ and $g(a,b) = h(a,b)$. Also $(a',b'), (a, \quad , h(a,b) \notin R(g)$, and $h(a',b') = h(a,b)$ jointly imply that $\quad a'$ and $b = b'$.

$[d_1^7]$ = Complete extension chain of halfgroupoids $\{\mathcal{H}_i \mid i \in N\}$ is an extension chain of halfgroupoids $\{\mathcal{H}_i \mid i \in N\}$ such that for every i $\mathcal{H}_{i+1}$ is a complete extension of $\mathcal{H}_i$.

Table IV (Continued)

$[d_2^7]$ = Halfgroupoid $\mathcal{K}$ is free over its subhalfgroupoid $\mathcal{J}$ provided every homomorphism of $\mathcal{J}$ into a halfgroupoid $\mathcal{H}$ can be extended to a homomorphism of halfgroupoid $\mathcal{K}$ into $\mathcal{H}$.

$[d_3^7]$ = Length of a finite divisor chain $\{a_k | k \in N\}$ over halfgroupoid $\mathcal{J}$ is the integer $n = \min k$ where $k$ is such that for every $p \in N$ $a_{k+p} = a_k$.

$[d_4^7]$ = Maximal extension chain of a halfgroupoid $(J,g)$ in a halfgroupoid $(H,h)$ is an extension chain $\{(J_i,g_i) | i \in N\}$ such that $(J_o,g_o) = (J,g)$, and $i \in N$ implies that (1) $J_i \subset H$ and (2) for $x,y \in J_i$ $g_{i+1}(x,y) = z$ whenever $h(x,y) = z$.

$[d_6^7]$ = Open extension chain of halfgroupoids is an extension chain of halfgroupoids $\{\mathcal{K}_i | i \in N\}$ such that for every integer $i$ $\mathcal{K}_{i+1}$ is an open extension of $\mathcal{K}_i$.

$[d_1^8]$ = Subhalfgroupoid of halfgroupoid generated by a subhalfgroupoid. Let $\mathcal{K}$ be a halfgroupoid and $\mathcal{J} \subset \mathcal{K}$. Let $\{\mathcal{J}_i | i \in N\}$ be a maximal extension chain of $\mathcal{J}$ in $\mathcal{K}$ and let $\mathcal{H} = \bigcup \mathcal{J}_i$ then $\mathcal{H}$ is subhalfgroupoid of $\mathcal{K}$ generated by $\mathcal{J}$.

$[d_1^9]$ = Halfgroupoid $\mathcal{K}$ is freely generated by halfgroupoid $\mathcal{J}$ if $\mathcal{J}$ generates $\mathcal{K}$ and $\mathcal{K}$ is free over $\mathcal{J}$.

$[d_1^{10}]$ = Free basis of a halfgroupoid $(H,h)$ is a subset $B \subset H$ such that $B$ freely generates $(H,h)$.

$[d_2^{10}]$ = Free product of a disjoint collection of halfgroupoids $\{\mathcal{K}_\alpha | \alpha \in A\}$ is a groupoid $\mathcal{J}$ freely generated by $\bigcup_{\alpha \in A} \mathcal{K}_\alpha$.

$[d_3^{10}]$ = Generalized free product of a compatible collection of halfgroupoids $\{\mathcal{K}_\alpha | \alpha \in A\}$ is a groupoid $\mathcal{J}$ freely generated by $\bigcup_{\alpha \in A} \mathcal{K}_\alpha$.

$[d_1^{11}]$ = Halfgroupoid is free if it has a free basis.

$[d_2^{11}]$ = Generalized free product of disjoint collection of groupoids with amalgamated subgroupoids $\{\mathcal{J}_\alpha, \mathcal{K}_{\alpha\beta}, \mathcal{K}_\alpha, \varphi_\alpha | \alpha,\beta \in A\}$ is a generalized free product of compatible collection of groupoids: $\mathcal{K} = \Pi^* \mathcal{K}_\alpha$.

Our wording is in most cases not an exact copy of that in [9].

Table V

| | | |
|---|---|---|
| $d_1^1$ | $d_3^0$ $d_{26}^0$ | $d_7^0$ $d_9^0$ $d_{11}^0$ $d_{20}^0$ $d_{23}^0$ $d_{25}^0$ |
| $d_1^2$ | $d_1^1$ | $d_7^0$ $d_9^0$ $d_{11}^0$ $d_{20}^0$ $d_{23}^0$ |
| $d_2^2$ | $d_1^1$ | $d_7^0$ $d_{11}^0$ $d_{20}^0$ $d_{23}^0$ |
| $d_3^2$ | $d_{26}^0$ | $d_1^1$ $d_7^0$ |
| $d_1^3$ | $d_{20}^0$ | $d_1^2$ $d_{11}^0$ $d_{23}^0$ |
| $d_1^4$ | $d_1^3$ $d_{25}^0$ | $d_3^2$ $d_7^0$ $d_9^0$ $d_{13}^0$ $d_{15}^0$ $d_{20}^0$ |
| $d_2^4$ | $d_1^3$ $d_{25}^0$ | $d_7^0$ $d_8^0$ $d_9^0$ $d_{13}^0$ $d_{15}^0$ $d_{18}^0$ |
| $d_3^4$ | $d_1^3$ $d_7^0$ | $d_9^0$ $d_{25}^0$ |
| $d_4^4$ | $d_1^3$ | $d_3^0$ $d_3^2$ $d_9^0$ $d_{25}^0$ |
| $d_5^4$ | $d_1^1$ | $d_1^3$ $d_7^0$ $d_9^0$ $d_{25}^0$ |
| $d_6^4$ | $d_8^0$ | $d_1^3$ $d_{25}^0$ |
| $d_7^4$ | $d_1^3$ | $d_7^0$ $d_9^0$ $d_{19}^0$ $d_{20}^0$ $d_{25}^0$ |
| $d_1^5$ | $d_1^3$ | $d_8^0$ $d_8^4$ |
| $d_2^5$ | $d_{24}^0$ | $d_1^3$ $d_3^4$ $d_7^0$ $d_{14}^0$ $d_{27}^0$ |
| $d_3^5$ | $d_8^4$ | $d_1^3$ |
| $d_4^5$ | $d_1^3$ | $d_3^0$ $d_9^2$ $d_7^4$ $d_{26}^0$ |
| $d_5^5$ | $d_1^3$ | $d_8^4$ $d_{10}^0$ |
| $d_6^5$ | $d_8^4$ | $d_1^1$ $d_9^0$ |
| $d_7^5$ | $d_1^4$ | $d_7^0$ $d_9^0$ $d_{18}^0$ $d_{18}^0$ |
| $d_8^5$ | $d_4^4$ | $d_1^1$ $d_{26}^0$ |
| $d_9^5$ | $d_7^0$ | $d_3^4$ $d_7^0$ $d_{25}^0$ |
| $d_{10}^5$ | $d_7^4$ | $d_1^3$ $d_4^4$ |

Table V (Continued)

| | | |
|---|---|---|
| $d_{11}^{5}$ | $d_{7}^{4}$ | $d_{3}^{2}\ d_{7}^{0}\ d_{20}^{0}\ d_{25}^{0}$ |
| $d_{12}^{5}$ | $d_{1}^{3}$ | $d_{1}^{4}\ d_{7}^{0}\ d_{9}^{0}\ d_{20}^{0}\ d_{25}^{0}\ d_{29}^{0}$ |
| $d_{1}^{6}$ | $d_{8}^{5}$ | |
| $d_{2}^{6}$ | $d_{4}^{5}$ | $d_{1}^{3}\ d_{3}^{0}\ d_{3}^{2}\ d_{25}^{0}\ d_{26}^{0}$ |
| $d_{3}^{6}$ | $d_{2}^{4}$ | $d_{1}^{1}\ d_{1}^{4}\ d_{4}^{4}\ d_{7}^{4}\ d_{8}^{5}\ d_{7}^{0}\ d_{9}^{0}\ d_{25}^{0}$ |
| $d_{4}^{6}$ | $d_{24}^{0}$ | $d_{1}^{3}\ d_{4}^{5}\ d_{7}^{0}\ d_{14}^{0}\ d_{27}^{0}$ |
| $d_{6}^{6}$ | $d_{2}^{5}$ | $d_{1}^{0}\ d_{1}^{3}\ d_{9}^{0}\ d_{14}^{0}\ d_{25}^{0}$ |
| $d_{8}^{6}$ | $d_{1}^{3}$ | $d_{7}^{4}\ d_{8}^{5}$ |
| $d_{7}^{6}$ | $d_{5}^{4}$ | $d_{1}^{1}\ d_{7}^{0}\ d_{9}^{0}\ d_{11}^{5}\ d_{25}^{0}$ |
| $d_{8}^{6}$ | $d_{1}^{3}$ | $d_{3}^{0}\ d_{3}^{2}\ d_{7}^{0}\ d_{7}^{5}\ d_{9}^{0}\ d_{15}^{0}\ d_{21}^{0}\ d_{25}^{0}$ |
| $d_{9}^{6}$ | $d_{4}^{5}$ | $d_{8}^{2}\ d_{7}^{0}\ d_{9}^{0}\ d_{17}^{0}\ d_{20}^{0}\ d_{25}^{0}$ |
| $d_{1}^{7}$ | $d_{4}^{6}$ | $d_{8}^{6}\ d_{14}^{0}\ d_{24}^{0}\ d_{27}^{0}$ |
| $d_{2}^{7}$ | $d_{1}^{3}$ | $d_{8}^{4}\ d_{7}^{4}\ d_{7}^{6}$ |
| $d_{3}^{7}$ | $d_{14}^{0}$ | $d_{1}^{3}\ d_{8}^{6}\ d_{7}^{0}\ d_{9}^{0}\ d_{1}^{0}\ d_{18}^{0}\ d_{27}^{0}$ |
| $d_{4}^{7}$ | $d_{4}^{6}$ | $d_{1}^{3}\ d_{7}^{0}\ d_{9}^{0}\ d_{14}^{0}\ d_{26}^{0}\ d_{27}^{0}$ |
| $d_{5}^{7}$ | $d_{4}^{6}$ | $d_{1}^{3}\ d_{8}^{6}\ d_{14}^{0}\ d_{27}^{0}$ |
| $d_{1}^{8}$ | $d_{7}^{4}$ | $d_{1}^{3}\ d_{4}^{7}\ d_{12}^{5}$ |
| $d_{1}^{9}$ | $d_{8}^{7}$ | $d_{1}^{8}$ |
| $d_{1}^{10}$ | $d_{26}^{0}$ | $d_{1}^{3}\ d_{1}^{9}\ d_{25}^{0}$ |
| $d_{8}^{10}$ | $d_{4}^{4}$ | $d_{1}^{9}\ d_{8}^{4}\ d_{12}^{5}$ |
| $d_{3}^{10}$ | $d_{4}^{4}$ | $d_{1}^{4}\ d_{1}^{9}\ d_{12}^{5}$ |
| $d_{1}^{11}$ | $d_{1}^{3}$ | $d_{1}^{10}$ |
| $d_{2}^{11}$ | $d^{10}$ | $d_{3}^{8}$ |

Table V consists of the addresses (column 1) $d_i^k$ of all the elements in $D_{12}$ and it gives the corresponding sets ${}^kD_i$ and $D_i^k$ in columns 2 and 3, respectively. In our example we choose the same reference for all the elements. For instance, we may write $\{d_1^3\}$ = p.1 [9], where [9] stands for the complete reference as given at the end of this report. Assignment of some integer value, say 100, to every rescission index completes the description of the set $D_{12}$ (with ${}_{12}D = \phi$) of our information.

Our example contains nineteen lemmas and theorems. Only one of them has a name, namely, $\tilde{t}_1^{12}$ = Free Representation Theorem. Therefore we omit the corresponding list of addresses and names. The addresses of the elements of the types s and t are contained in Table VI, which gives the references $\{s_i^k\}$ and $\{t_i^k\}$ of these elements. The sets $(s_i^k)$, $(t_i^k)$, $(\tilde{s}_i^k)$, and $(\tilde{t}_i^k)$ of our example are all empty.

Table VI

$\{s_1^9\}$ = p.2, [9]

$\{t_1^9\}$ = Lemma 1.1, p.2, [9]

$\{t_2^9\}$ = Lemma 1.2, p.3, [9]

$\{t_3^9\}$ = Theorem 1.8, p.7, [9]

$\{s_1^{10}\}$ = Lemma 1.4, p.4, [9]

$\{t_1^{10}\}$ = Theorem 1.1, p.4, [9]

$\{t_2^{10}\}$ = Theorem 1.1, p.4, [9]

$\{s_1^{12}\}$ = Lemma 1.3, p.3, [9]

$\{s_2^{12}\}$ = Lemma 1.5, p.6, [9]

$\{s_3^{12}\}$ = Theorem 1.6, p.6, [9]

$\{t_1^{12}\}$ = Theorem 1.2, p.5, [9]

$\{t_2^{12}\}$ = Theorem 1.3, p.5, [9]

Table VI (Continued)

$\{t_3^{12}\}$ = Lemma 1.5, p.6, [9]

$\{t_4^{12}\}$ = Lemma 1.5, p.6, [9]

$\{t_5^{12}\}$ = Theorem 1.4, p.6, [9]

$\{t_6^{12}\}$ = Lemma 1.6, p.6, [9]

$\{t_7^{12}\}$ = Theorem 1.5, p.6, [9]

$\{t_8^{12}\}$ = Theorem 1.7, p.6, [9]

$\{t_9^{12}\}$ = p.8, [9]

The second column of Table VII shows the sets ${}^kS_i$ and ${}^kT_i$ that correspond to addresses in the first column. The third column contains elements of the sets $S_i^k$ and $T_i^k$.

Table VII

| | | |
|---|---|---|
| $s_1^9$ | $d_1^3$ $d_1^8$ $d_7^4$ | $d_{11}^5$ $d_7^4$ |
| $s_1^{10}$ | $d_1^9$ | $d_9^5$ $d_7^0$ $d_8^0$ $d_9^0$ $d_2^5$ $d_{10}^0$ $d_5^6$ |
| $s_1^{12}$ | $d_4^5$ | $d_1^{11}$ $d_6^7$ |
| $s_2^{12}$ | $d_1^{11}$ | $d_{25}^0$ $d_7^0$ $d_8^5$ $d_{20}^0$ $d_3^0$ $d_9^0$ $d_2^5$ $d_5^8$ |
| $s_3^{12}$ | $d_1^{11}$ | $d_9^{10}$ $d_6^4$ $d_9^0$ $d_{14}^0$ $d_1^{11}$ |
| $t_1^8$ | $d_1^8$ | |
| $t_2^9$ | $d_1^3$ $d_1^8$ $d_6^4$ | $d_1^3$ $d_6^4$ $d_7^8$ |
| $t_3^8$ | $d_6^5$ | $d_1^8$ $d_6^8$ $d_7^0$ |
| $t_1^{10}$ | $d_1^3$ | $d_1^9$ $d_4^4$ |
| $t_2^{10}$ | $d_1^3$ $d_1^9$ | $d_1^3$ $d_6^5$ $d_4^4$ $d_{12}^0$ |
| $t_1^{12}$ | $d_1^3$ | $d_1^{11}$ $d_6^5$ |

Table VII (Continued)

| | | |
|---|---|---|
| $t_2^{12}$ | $d_4^4\ d_6^5\ d_1^{11}$ | $d_{11}^0\ d_{23}^0\ d_4^4\ d_6^5\ d_1^{11}\ d_7^4\ d_8^5$ $d_{12}^5\ d_9^0\ d_1^1$ |
| $t_3^{12}$ | $d_1^{11}$ | $d_{25}^0\ d_9^5\ d_1^{10}$ |
| $t_4^{12}$ | $d_1^{11}\ d_1^{10}$ | $d_{25}^0\ d_9^5$ |
| $t_5^{12}$ | $d_1^{11}\ d_{11}^5$ | $d_1^{11}$ |
| $t_6^{12}$ | $d_{11}^5\ d_4^0\ d_{18}^0\ d_1^8\ d_{22}^0\ d_1^{11}\ d_{28}^0$ | $d_1^{10}$ |
| $t_7^{12}$ | $d_1^{11}\ d_4^4$ | $d_1^{11}\ d_4^4\ d_7^4\ d_6^0\ d_8^4$ |
| $t_8^{12}$ | $d_4^4\ d_1^9\ d_7^4\ d_{25}^0\ d_9^5\ d_9^0\ d_{15}^0\ d_8^0$ | $d_4^4\ d_1^9\ d_{15}^0\ d_8^0\ d_{18}^0\ d_3^{10}\ d_1^{11}$ $d_1^{10}$ |
| $t_9^{12}$ | $d_3^3\ d_{25}^0\ d_1^1\ d_8^5\ d_2^2\ d_7^0\ d_{13}^0$ $d_{18}^0\ d_9^0\ d_{28}^0\ d_{11}^0\ d_{23}^0\ d_{26}^0$ | $d_1^8\ d_4^4\ d_8^5\ d_2^{11}$ |

Next we give the verbal statements of these elements

$[s_1^9]$. A subhalfgroupoid $\mathcal{Q}$ of a halfgroupoid $\mathcal{K}$ generates $\mathcal{K}$ *if and only if* every closed subhalfgroupoid of $\mathcal{K}$ containing $\mathcal{Q}$ is equal to $\mathcal{K}$.

$[t_1^9]$. If a subhalfgroupoid $\mathcal{Q}$ of a halfgroupoid $\mathcal{K}$ generates a subhalfgroupoid $\mathcal{K}$ of $\mathcal{K}$ and $\mathcal{K}$ generates $\mathcal{K}$ *then* $\mathcal{Q}$ generates $\mathcal{K}$.

$[t_2^9]$. If $\mathcal{Q}$ is a generating subhalfgroupoid of a halfgroupoid $\mathcal{K}$ and if $\varphi$ is a homomorphism of $\mathcal{Q}$ into a halfgroupoid $\mathcal{K}$, *then* $\varphi$ can be extended in at most one way to a homomorphism $\theta$ of $\mathcal{K}$ into $\mathcal{K}$.

$[t_3^9]$. If $\mathcal{Q}$ is a finite halfgroupoid, *then* there exists a halfgroupoid $\mathcal{K}$ generated by one element of $\mathcal{Q}$ and imbeddable in $\mathcal{K}$.

$[s_1^{10}]$. A halfgroupoid $\mathcal{K}$ is freely generated by a subhalfgroupoid $\mathcal{Q}$ *if and only if* the following conditions hold

(i) If $a$ is a prime in $\mathcal{Q}$, then $a$ is prime in $\mathcal{K}$.

(ii) If $a \in \mathcal{K}$, $a \notin \mathcal{Q}$, then $a = bc$ in $\mathcal{K}$ for one and only one ordered pair $b,c \in \mathcal{K}$.

(iii) Every divisor chain is either finite or finite over $\mathcal{Q}$.

$[t_1^{10}]$. If $\mathcal{Q}$ is a halfgroupoid, then there exists a groupoid $\mathcal{K}$ freely generated by $\mathcal{Q}$.

$[t_2^{10}]$. If a halfgroupoid $\mathcal{Q}$ freely generates groupoids $\mathcal{K}$ and $\mathcal{K}$, then there exists an isomorphism $\varphi$ of $\mathcal{K}$ onto $\mathcal{K}$ such that $\varphi$ is an identity mapping on $\mathcal{Q}$.

$[s_1^{12}]$. If a halfgroupoid $\mathcal{K}$ is an extension of a halfgroupoid $\mathcal{Q}$, then $\mathcal{K}$ is free over $\mathcal{Q}$ if and only if $\mathcal{K}$ is an open extension of $\mathcal{Q}$.

$[s_2^{12}]$. A halfgroupoid $\mathcal{K}$ is free if and only if the following conditions hold:

(i) If $a \in \mathcal{K}$ and if $a$ is not prime in $\mathcal{K}$, then $a = bc$ in $\mathcal{K}$ for one and only one ordered pair $b,c \in \mathcal{K}$.

(ii) Every divisor chain in $\mathcal{K}$ is finite.

$[s_3^{12}]$. A groupoid $\mathcal{Q}$ is free if and only if $\mathcal{Q}$ is a free product of free groupoids of rank one.

$[t_1^{12}]$. If $\mathcal{Q}$ is a groupoid, then there exists a free groupoid $\mathcal{K}$ isomorphic with $\mathcal{Q}$.

$[t_2^{12}]$. If $\theta$ is a homomorphism of the groupoid $\mathcal{Q}$ onto a free groupoid $\mathcal{J}$, then there exist a subgroupoid $\mathcal{K}$ of $\mathcal{Q}$ and an isomorphism $\varphi$ of $\mathcal{J}$ onto $\mathcal{K}$ such that $\theta(\varphi(x)) = x$ for every $x \in \mathcal{J}$.

$[t_3^{12}]$. If $\mathcal{K}$ is a free halfgroupoid, then the set of all primes in $\mathcal{K}$ is a free basis of $\mathcal{K}$.

$[t_4^{12}]$. If $\mathcal{K}$ is a free halfgroupoid and B is a free basis of $\mathcal{K}$, then B is the set of all primes in $\mathcal{K}$.

$[t_5^{12}]$. If $\mathcal{K}$ is a subhalfgroupoid of a free halfgroupoid $\mathcal{Q}$, then $\mathcal{K}$ is free.

$[t_6^{12}]$. Suppose $\mathcal{K} = (H,f)$ is a free halfgroupoid, $B \subset H$, $B \neq \emptyset$ and B generates a closed subhalfgroupoid $\mathcal{K}_1$ of $\mathcal{K}$. If no proper subset of B generates $\mathcal{K}_1$, then B is a free basis of $\mathcal{K}_1$.

$[t_7^{12}]$. If $\underline{\mathcal{K}}$ is a free groupoid, then there exists a subgroupoid $\underline{\mathcal{L}}$ of $\underline{\mathcal{K}}$ such that $\underline{\mathcal{L}}$ is free and of countable order.

$[t_8^{12}]$. Let the groupoid $\underline{\mathcal{L}}$ be freely generated by a subhalfgroupoid $\underline{\mathcal{F}} = (\underline{F},\underline{f})$ and let $\underline{\mathcal{K}} = (H,h)$ be a subgroupoid of $\underline{\mathcal{L}}$. Let $\underline{P}$ be the set of all primes of $\underline{\mathcal{K}}$ which are not in $\underline{F} \cap \underline{H}$. Then one of the following holds:

(i) $\underline{P}$ is empty, $\underline{F} \cap \underline{H}$ is not empty and $\underline{\mathcal{K}}$ is freely generated by $\underline{F} \cap \underline{H}$; or

(ii) $\underline{F} \cap \underline{H}$ is empty, $\underline{P}$ is not empty and $\underline{P}$ is a free basis of $\underline{\mathcal{K}}$; or

(iii) Neither $\underline{P}$ nor $\underline{F} \cap \underline{H}$ is empty and $\underline{\mathcal{K}}$ is a generalized free product of $\underline{\mathcal{L}}$ and $\underline{\mathcal{M}}$, where $\underline{\mathcal{L}}$ is a free subgroupoid of $\underline{\mathcal{K}}$ with free basis $\underline{P}$ and $\underline{\mathcal{M}}$ is a subgroupoid of $\underline{\mathcal{K}}$ freely generated by $\underline{F} \cap \underline{H}$.

$[t_9^{12}]$. Let $\{\underline{\mathcal{K}}_\alpha = (H_\alpha,f_\alpha) \mid \alpha \in A\}$ be a disjoint collection of groupoids with amalgamated subgroupoids $\underline{K}(\alpha,\beta)$. Let $\underline{\theta}_{\alpha\beta}$ be an isomorphism of $\underline{K}(\alpha,\beta)$ onto $\underline{K}(\beta,\alpha)$. Suppose that $\underline{\theta}_{\alpha\beta}^{-1} = \underline{\theta}_{\beta\alpha}$ and that $\underline{x} \in \underline{K}(\alpha,\beta)$ together with $\underline{\theta}_{\alpha\beta}(x) \in \underline{K}(\beta,\gamma)$ imply that $\underline{x} \in \underline{K}(\alpha,\gamma)$ and $\underline{\theta}_{\alpha\gamma}(x) = \theta_{\beta\gamma}(\theta_{\alpha\beta}(x))$. Let $\underline{H} = \cup H_\alpha$, $\underline{\mathcal{K}} = \cup \underline{\mathcal{K}}_\alpha = (H,f)$ and for $\underline{a},\underline{b} \in \underline{H}$ let $\underline{a} \equiv b$ whenever either $a = b$ or there exist $\underline{\alpha},\underline{\beta}$ such that $\underline{\alpha} \neq \underline{\beta}$, $\underline{a} \in \underline{K}(\alpha,\beta)$, $\underline{b} \in \underline{K}(\beta,\alpha)$, and $\underline{\theta}_{\alpha\beta}(a) = b$. Let $\underline{S}_a$ be a subset of $\underline{H}$ such that for every $\underline{b} \in S_a$, $\underline{a} \equiv b$ and for every $\underline{c} \in \underline{H}$ with $\underline{c} \equiv a$, $\underline{c} \in \underline{S}_a$. Let $\underline{S} = \{S_a \mid a \in H\}$ and for $\underline{x}$, $\underline{y}$, $\underline{z} \in \underline{S}$ let $\underline{g}(x,y) = z$ provided for some $\underline{a} \in x$, $b \in y$, and $c \in z$ we have $f(a,b) = c$. Let $\underline{\mathcal{K}'} = (S,g)$ and $\underline{\mathcal{K}''}$ be a groupoid generated by the halfgroupoid $\underline{\mathcal{K}'}$. Then $\underline{\mathcal{K}''}$ is isomorphic to a generalized free product of a disjoint collection of groupoids with amalgamated subgroupoids.

Relations between the elements as defined in Section II can be easily constructed from Tables V and VII. Also k-chains of elements can be obtained by consulting these Tables. A k-chain of relations can be written down by first constructing a corresponding k-chain of elements and then by taking the natural relations from the lists of verbal statements of the elements. Also the matrices $M_\chi$ that represent the

structure of $J_k = I_k \setminus I_{k-1}$ can be easily constructed from Tables VI and VII. For instance, matrix $M_1$ is a 29 × 1 matrix with elements in rows 3, 7, 9, 11, 20, 23, 25, 26 equal to one and with zeros in the remaining positions. The matrix $M_{12}$ is 74 × 12. The row $d_1^{1a}$ of Table VII indicates that the first column of $M_{12}$ will have a one in rows 34, 49, and 73. The remaining elements of this column are zeros. The other elements can be easily determined by examining Table VII and by obtaining sequential numbers of elements $d_i^k$ from the combined sequence of first columns in Tables I and V.

## IV. RETRIEVAL OF INFORMATION

Our information system consists of elements and certain relations induced by the structure of the elements. Hence, these elements, their components and their relations are what we can retrieve. Of course, the choice of the elements was made in order to provide a retrieval of information that would satisfy most of the needs of a researcher in the field. We hope that we have succeeded to some degree. In fact, the starting point for the choice of our definition was a collection of likely questions that one may want to ask.

In the present section we try to describe these questions in general terms or as classes of questions. Next we attempt to construct the rules for obtaining the answers by retrieval of elements of our information, their components, and their relations. Of course, we assume that the system can be interrogated by formulating new requests for information according to responses obtained to previous questions.

We are only interested in principles of operation. Therefore in our discussion it does not matter how retrieved information is displayed. We may assume that the system can provide a temporary display of retrieved information for scanning purposes as well as produce some sort of a permanent copy.

It seems that answers to the following classes of questions should satisfy many of the needs of the user of information in a discipline of mathematics.

($\alpha$) How is a term c defined?

($\beta$) Are terms b and c synonyms?

($\gamma$) What are special cases of concept c?

($\delta$) What are generalizations of concept c?

($\varepsilon$) Does there exist a concept d such that b and c are special cases of d?

($\eta$) Does there exist a concept d such that both b and c are generalizations of d?

($\xi$) Are statements q and r equivalent?

($\theta$) Does q imply r?

($\iota$) What is an exact wording of a theorem on subject q, if any?

($\varkappa$) Under what conditions does q imply r?

($\lambda$) Does there exist a theorem that relates q and r?

($\mu$) What are generalizations of theorem t?

($\nu$) What are special cases of theorem t?

Of course, question ($\theta$) and many other questions on this list can be asked about postulates or conjectures as well as about theorems. Also, we may inquire about a symmetric theorem on our subject, not just any theorem, i.e. we may want to know both necessary and sufficient conditions or just sufficient conditions. Therefore we assume that the user may choose to specify which of the sets $P_N$, $C_N$, $S_N$, and $T_N$ should be examined.

The user may not be familiar with the dictionary of the information system. Therefore he may want to begin his search with certain preliminary questions that are equivalent to consultation of some sort of "thesaurus". In our case a "thesaurus" consists of the main dictionary of the system as well as of $^{k}D$-, $^{k}P$-, $^{k}C$-, $^{k}S$-, and $^{k}T$ - dictionaries with k=1,2,...,N, as described in Section II above. We

allow three classes of preliminary questions, that we call "dictionary questions".

(1) Is the term c in the dictionary? This question may include a request for synonyms and the main term of c.

(2) Which terms of the System I appear in the definition of the term c?

(3) Which terms are in the system? This question may be about any of the following collections of terms:

($3^{i}$) All the terms in the system.

($3^{ii}$) All the terms in $D_N = D$, i.e. all the terms that represent concepts as opposed to names of theorems and so on.

($3^{iii}$) All the terms contained in $J_k$ for some k.

($3^{iv}$) All the terms in $D_k$.

Of course, the same procedure that is used to examine the terms in $D_k$ and D can be applied to scan the terms in $A_k$ and $A_N$, for A = P, C, T, or S.

Questions ($\alpha$) and ($\beta$) are questions on elements or their components. We call these "element questions".

The remaining questions ($\gamma$) - ($\nu$) are on relations of elements. Questions ($\gamma$) - ($\eta$) are about relations between concepts. Consequently, they do not involve any semantics and hence can be answered very easily. The remaining questions involve statements q and r. These statements may consist of a single term. For instance, the question: "which topological spaces are metrizable" is of class ($\chi$) with q = topological space and r = metrizable, i.e. both q and r consisting of single concepts. However, frequently q and r will be more complex. In our opinion no increase in storage capacity and in processing speed of present electronic devices and no sophistication of accompanying software will produce a system able to read and interpret the semantics

of a living language with all its variations from person to person and with its continuous growth and changes. One can build an automaton to handle only a rigid, fixed, and limited language. Therefore automatic answering of questions with semantics can be accomplished only for a very rigid language. On the other hand, we should not expect that every information seeker will learn this inflexible baby talk of a computer. Therefore an automatic answering system should not depend on semantics. We propose to limit semantics to a selection of the class of a question. Since our list of classes contains only thirteen items, ($\alpha$) through ($\nu$) above, it is not too much to ask the user to select the class to which, in his opinion, his question belongs. The rest of the semantics that is expressed by the linguistic form of statements q and r is discarded by constructing two sets E and F that consist of concepts in our main dictionary. The set E contains the terms that appear in the statement q and the set F consists of terms in r. We may have a very large number of different sets of the type E and F, but an automatic system would have no difficulty in recognizing when such sets are identical, when an inclusion relation holds, and when it does not hold.

Of course, replacement of the statements q and r by the sets E and F will result in retrieval of irrelevant information. For instance, if we ask if q implies r and if we specify the class of question ($\theta$) and at the same time replace q and r by E and F respectively, we may retrieve a theorem "q' implies r'", where q' and r' are stated in the terms contained in E and F respectively, and yet the retrieved theorem may have no relation to our original question. However, it seems that in a restricted field of science the incidence of such irrelevant responses will not be very high, and the user will not have to waste much time in sorting out irrelevant answers from relevant information.

We assume that the system constructs the sets E and F automatically. A question of the user should be fed into a system in its original semantic form. However the user should specify which part of the question contains the statement q and which part is a phrasing of

the statement r. The system should scan the question and construct the sets E and F by including those terms that are in the main dictionary. We assume that our dictionary includes words that are synonyms in the technical jargon of the particular discipline, such as "metrizable" and "metrizability". These may be our primary synonyms. Furthermore, the system should be able to accept secondary synonyms, that - by the choice of the designer of the system - may include certain semantic variants of primary synonyms such as plural or possessive form, past tense, or conjunctive mode. It is still possible that the user will use a form that is not in the main dictionary. Such a term would be omitted in constructing E and F. Hence the user should either carefully check the vocabulary or else select judiciously the class to which his question belongs. For instance, even with some terms omitted from E or F a correct answer may result if instead of class ($\theta$) the class ($\varkappa$) is selected.

In our opinion, by successive choice of questions one can obtain the desired information, provided it is contained in the system.

Of course, we do not think that a system should be designed capable of answering questions about itself such as how complete it is or what is the range of applicability of the information. Nor should the system be intended to supply any evaluation of its contents such as complexity of the proofs of theorems, etc.

The form of some questions listed above may suggest that answers may be simply "yes" or "no". However, we assume that in many cases the answer is more complete. For instance, the answer to ($\xi$) may include a verbal statement of a particular theorem. Verbal statements need not be restricted to synonyms included in the system. For instance, the verbal statement $[s_1^9]$ of our example in the preceding section contains the words "closed halfgroupoid of $\underline{\mathcal{K}}$". Certainly, $\underline{\mathcal{K}}$ is not a part of our vocabulary. Its meaning follows from the preceding sequence of words in $[s_1^9]$. Besides a verbal statement, the user frequently may desire a reference, where he could find a proof of a theorem and also additional references.

We can express the questions listed above in terms of requests for elements of information, their components, and their relations. We will use this formulation in the next section to construct rules of operations that would retrieve the desired information. We begin with dictionary questions.

Question (1) is a search for a main term that corresponds to a chosen synonym or its variant. For this we must scan certain portions of our dictionaries. The scanning may be facilitated if we know the type of the term. For instance, we may want to search for a name of a theorem or for a concept. Again we may know that the concept is rather basic and hence most likely is an element of information of low level. Therefore, instead of searching in the $D_N$-dictionary we may choose to scan the corresponding portion of $D_k$-dictionary for some $k < N$. If no dictionary is specified, then the system should scan the main dictionary. If a match with the term c is found, the address $a_i^k$ and the corresponding main term are given. If no such match is found the answer is that the term is not in the information. The user may specify which components of the element should be displayed or printed. A request for the third and fourth component would answer question (2) as stated above. Similar search is performed in any other dictionary specified by the user.

Dictionary question (3) can be answered by a visual scanning of a specified portion of a selected dictionary. A dictionary is selected by indicating $D_k$-dictionary, $I_k$-dictionary, etc. If a dictionary is not specified then the main dictionary is used. The portion of the specified dictionary to be scanned is indicated by a set of, say, three letters such as, for instance, COL or DIS when it is desired to check whether the term "disjoint collection of halfgroupoids" or "collection of disjoint halfgroupoids" are in the information. (It happens that in our example the first is a concept or a term; the second is a phrase or a statement). After a dictionary and a beginning searching place are chosen, the system should display successive words in the dictionary beginning at the specified place. At any point the user may terminate

the search and may request the display of any desired component of the element that was displayed at the time of termination. Or else the user may start his search in a different dictionary or at another place of the same one.

The two element questions ($\alpha$) and ($\beta$) can be answered by first determining the address $a_i^k$ (or addresses $a_i^k$ and $b_j^m$ in the case of ($\beta$)). These addresses can be obtained as in question (1). Identity of both addresses would answer question ($\beta$) affirmatively. Different addresses would mean that b and c are not synonyms. The answer to question ($\alpha$) would be obtained by retrieving the verbal statement $[a_i^k]$. Additionally, one may desire to retrieve the reference $\{a_i^k\}$ or synonyms $(a_i^k)$, i.e. to retrieve the sixth or the second component of $a_i^k$.

The answer to question ($\gamma$) consists of the addresses of all the elements of the set D for which $c = {}^kA_i$. This answer, by the choice of the user, may contain addresses, or main terms, or verbal statements, or any combination of these. Usually $c \in {}_mD$ and it may be desirable to limit the search to elements that belong to ${}_kD$ for some k, or that belong to the union of some collection of such sets.

This procedure would recover any special cases of the concept c that are introduced by definitions. However c may be a generalization of a concept b by virtue of the semantic meaning of various definitions of lower level in terms of which both c and b are defined or by virtue of the meaning of terms that belong to $D_o$, i.e. the terms that are not defined in the information. The logical relation between b and c in this case can be established by a theorem whose proof may be very simple or very difficult. In any case a search for this type of special case is a search for theorems either of the form "c implies b" or "every c with the property f is a b". A search for this type of generalization is a question of some class from ($\iota$) to ($\mu$). Retrieving answers to these questions is described below.

The answer to question ($\delta$) is simply the third component ${}^{k}D_i$ of $d_i^k = c$. The user may choose to repeat the retrieval of the third component of the element contained in ${}^{k}D_i$ and so on. This would produce higher level generalizations if they exist. Again, in this way generalizations introduced by definitions can be retrieved. Other generalizations that are implied by the logical structure of the discipline are of the type discussed above in connection with question ($\gamma$). The remarks made there apply to this case also.

We obtain an answer to question ($\epsilon$) by constructing positive chains of elements from b and c, say, $\alpha_1,\alpha_2,\ldots$, and $\beta_1,\beta_2,\ldots$ with $b = \alpha_1$ and $c = \beta_1$. We add alternately one element to each chain and compare the added element with all the elements in the other chain that belong to information of the same level until we find a matching pair of elements or until both chains reach the maximum length. In the first case the matching element is the concept d. In the second case a concept d with the required properties does not exist.

Question ($\eta$) is analogous to ($\epsilon$). The answer can be obtained by constructing negative chains of elements beginning with b and c.

Question ($\xi$) can be directed at postulates, conjectures, or theorems. Hence the user should specify one of the sets ${}_{k}P$, $P_k$, ${}_{k}C$, $C_k$, ${}_{k}S$, or $S_k$ in which to search for a desired statement. After this the components ${}^{k}A_i$ and $A_i^k$ of the elements in the specified set are compared with E, the set of concepts in q. In the case of equality of the two sets, say, in the case when ${}^{k}A_i = E$ the set $A_i^k$ is compared with F, the set of concepts in r. If $F = A_i^k$ then the component $[a_i^k]$ is retrieved. The user decides whether this is a relevant statement or not and, accordingly, terminates his search or continues it further.

Question ($\theta$) can be handled in the same way as ($\xi$), except that the sets of the form ${}_{k}T$ and $T_k$ must be included in the search.

In the case of question ($\iota$) one selects ${}_kA$ or $A_k$ as in ($\xi$) and then the union of the components ${}^kA_i$ and $A_i^k$ of each element in the selected set is compared with the set E. When an element $a_i^k$ is found such that E is contained in ${}^kA_i \cup A_i^k$, the component $[a_i^k]$ is displayed or printed. The user decides whether this is a satisfactory answer to his question.

Question ($\chi$) requires a search for a theorem (or conjecture or postulate) with the consequent r and with an antecedant that contains q. In the search we compare F, the set of the concepts contained in r, with components ${}^kS_i$ and $S_i^k$ of the elements in ${}_kS$ or $S_k$ and with components $T_i^k$ of the elements in ${}_kT$ or $T_k$. When a component equal to F is found, one tests if the other of the two components ${}^kA_i$ or $A_i^k$ contains E. If it does the component $[a_i^k]$ is retrieved.

Handling of question ($\lambda$) is similar to that of ($\chi$) except that no distinction is made between sets of type S and type T. Furthermore, instead of searching for a component of an element that equals F one searches for a component that contains F as a subset.

In question ($\mu$) let q be the antecedant of t, and let r be the consequent. Generalization of t may be interpreted in the following ways: ($\mu_1$) number of conditions contained in q is reduced; ($\mu_2$) consequent r is changed to a consequent r' in such a way that r' applies to a larger class of entities than r; ($\mu_3$) antecedant q is changed to an antecedant q' that applies to a larger class of elements than q; ($\mu_4$) combination of ($\mu_1$) and ($\mu_2$); ($\mu_5$) combination of ($\mu_2$) and ($\mu_3$). Usually, a reduction of the number of conditions contained in q will increase the class of elements that satisfy the (reduced) set of conditions. Hence ($\mu_1$) and ($\mu_3$) may seem to be overlapping. However, we adopt the following interpretation of these cases that makes them disjoint. We assume that in the case ($\mu_1$) the change of q is the elimination of some conditions. This means that the set E is changed by removing some of its elements. In the case ($\mu_3$) the number of conditions is not changed, but some of the concepts in the conditions of q are replaced by their generalizations. According to this interpretation

the case ($\mu_1$) is analogous to ($\varkappa$). The case ($\mu_3$) can be handled in two steps. First, the user selects the terms of E that he may want to generalize and replaces them by generalizations of his choice. He may use the procedure described under ($\delta$) above to select such generalizations. Then he can choose the procedure in ($\xi$) - ($\lambda$). The case ($\mu_2$) is analogous to the case ($\mu_3$). Here again first some terms in F are replaced by their generalizations and the search of the type ($\xi$) - ($\lambda$) is applied. Similar combinations of previously described procedures would handle the cases ($\mu_4$) and ($\mu_5$). It should be noted that this procedure may not yield a desired result even when a certain generalization of a selected theorem is contained in the system. However, repeated requests for a certain type of generalization will lead to a desired theorem, or the user will find that no generalization in a chosen direction is contained in the system.

Question ($\nu$) is analogous to question ($\mu$). Here again we consider the following cases: ($\nu_1$) added conditions to the antecedant; ($\nu_2$) particularization of some concepts in r; ($\nu_3$) particularization of some concepts in q; ($\nu_4$) combination of ($\nu_1$) and ($\nu_3$); ($\nu_5$) combination of ($\nu_2$) and ($\nu_3$). Again, particularization of concepts is chosen by the user, probably, with the aid of the procedure described in ($\gamma$). The search is completed by choosing one of the procedures ($\xi$) - ($\lambda$).

## V. PRINCIPLES FOR IMPLEMENTATION

An element of information consists of eight pieces: its address $a_1^k$ and its seven components. Search procedures described in the preceding section require a direct access to any component and also to addresses contained in the third and the fourth components of an element. A system that satisfies these conditions could be constructed as a card catalog, although some search procedures with the cards would be very lengthy. Nevertheless we feel that a description of the principles in terms of the concrete concepts of a card file may be easier to read than an abstract discussion. Furthermore, access procedures for a card file can easily be interpreted as procedures carried out by electronic data

processing equipment. Later in this section we will discuss an idealized electronic system that, in our opinion, is feasible in principle, although, probably, nobody would consider building such a system, at least now.

Thus, we store each element of information on a card that contains the address of the element and all its components. The file of all the elements of information is arranged "lexicographically" with respect to their addresses. For this purpose we interpret the address $a_i^k$ as a three dimensional vector (a,k,i) with a = d, p, c, s, or t. It is convenient to assume the order of the values of a as just written. Of any two elements with the same first component in their addresses the one with the smaller k precedes the one with the larger k, and so on. We call this the main information file.

In order to answer our dictionary questions, (1) - (3) of the preceding section, we need a set of "dictionary files". The main dictionary would contain a card for each main term, and one for each primary and secondary synonym. These cards should be arranged in alphabetical order, and each should contain the corresponding address of the element. Most of the theorems, conjectures, and postulates have no names. Hence their main terms are identical with their addresses. Therefore these terms need not be included in the main dictionary or any other dictionary. To ease a search for a term we may construct any number of other card dictionaries that would implement the ${}_kA$- and $A_k$-dictionaries described in Section II. For instance, we may want to have two additional copies of the main dictionary cards. One copy would be divided into 5 x (N+1) parts according to the values of a and k and each part arranged in alphabetical order. Another copy could be divided into N+1 parts according to the value of k and each part arranged alphabetically. We may also choose to construct $I_k$-dictionaries as well as $A_k$-dictionaries for A = D, P, C, S, or T and for k=0,1,2,...,N. With these files representing our dictionaries we can answer all the dictionary questions of the preceding section. There is, of course, a trade-off between the number of the various dictionaries and the average duration of search for an answer to dictionary questions.

An element question, ($\alpha$) or ($\beta$) of the preceding section, can be answered in two steps: first one finds the address of a chosen element in the main dictionary or in any other dictionary, and then one selects the corresponding card from the information file.

Question ($\gamma$) can be answered easily if we have a ${}^{k}D_i$-file, i.e. a file of element cards of the type $d_i^k$ arranged lexicographically with respect to the address contained in ${}^{k}D_i$. Those element cards that have the same address in ${}^{k}D_i$ are arranged among themselves according to their main address. Now the answer to question ($\gamma$) can be obtained by first finding the address $d_j^m$ of c in a dictionary and then selecting the part of ${}^{k}D_i$-file with the address $d_j^m$ in $D_i^k$ and with k in the main address such that the inequalities $m > k > m-\ell$ ($\ell$ is an integer less or equal m) are satisfied. The selected cards will contain all the information that is needed for question ($\gamma$).

Question ($\delta$) can be answered by selecting the address in the third component of $d_j^m$, say, $d_{j_1}^{m_1}$, then by selecting the address in the third component of $d_{j_1}^{m_1}$, etc. until the desired level of information is reached. This selection is conveniently carried out by using the main information file.

Positive chains of elements that are required to answer the question ($\epsilon$) can be constructed by repeated application of the procedure described in ($\delta$) above.

Negative chains of elements as required in question ($\eta$) can be constructed by repeated application of the rules in ($\gamma$) above. Each time that this procedure produces more than one element we obtain branching of a negative chain. The elements in all the branches of the negative chains beginning with the element b must be compared with respective elements in the branches of negative chains beginning with c.

In order to answer questions ($\zeta$) - ($\nu$) we construct four additional files of the element cards, namely, a P-file, a C-file, an S-file, and a T-file. Each of these files consists of code-punched cards as described

on pp. 11-12 of [1 ]. The addresses of the elements in ${}^{k}A_i$ and $A_i^k$ are represented by the punched code in such a way that an insertion of a tumbler in a proper hole of the file permits us to retrieve all the cards containing a specified address. Answers to questions ($\zeta$) - ($\nu$) are obtained by testing whether a specified set E or F is a subset (proper or improper) of ${}^{k}A_i$ or $A_i^k$. Thus, with the aid of a tumbler one can first retrieve all the cards that contain the address of the first element in E. Next the same procedure is applied to the retrieved partial deck and the cards that contain the address of the second element of E are pulled out. A repetition of this procedure for the other elements of E would produce the desired collection of cards.

The set D may be very large, and it may be impossible to provide this type of punch-code for the address of every element in D. However, in principle, one can assume a sufficiently long card or the holes and notches sufficiently small to permit accommodation of all the necessary addresses.

The aspect just described brings out an important point of this system, namely the need to scan all the elements (cards) that contain certain addresses. We may say that all the cards containing a specified address must be connected in some sense. In order to obtain answers to questions ($\gamma$) - ($\eta$) we provide this "connectedness" by placing the corresponding cards consecutively in the ${}^{k}D_i$-file. For the remaining cases we achieve "connectedness" by using a tumbler together with coded holes and notches.

Just as in our human memory a concept such as, say, "groupoid" is the same in whatever context of other concepts it appears, so also a term in the information system should remain a single entity instead of a multiplicity of duplicates some of which can be changed or deleted without any effect on the remaining duplicates. We have a number of relations such as the relation between a groupoid and its base or its subgroupoid, but there is only one concept of groupoid. Hence, an information system should contain one single copy of each element and a

number of relations represented by proper connections between the elements. We may describe these connections in terms of some sort of conductors and, thus, avoid multiplicities of individual concepts. Thus, we assume that each element $a_j^m$ is connected to each set ${}^kA_i$ or $A_i^k$ of the element $a_i^k$, provided the concept represented by $a_j^m$ appears in $[a_i^k]$. With such connections present the requests of classes ($\alpha$) to ($\nu$) can be answered without searching or scanning the irrelevant elements of the system.

Such a system can be initially constructed and also expanded by building all seven components of an element at one time. We recall that our information system consists of two types of elements. Elements of level zero have only two components, $\tilde{d}_i^o$ and $(d_i^o)$. When a new element of this type is added to the system, its components (they consist of words) are "read" into the system and after this the respective dictionaries are automatically updated. First, the ${}_oD$-dictionary is updated by inserting the main term of the new element in its proper position according to alphabetical order. Let $d_{i_1}^o$ and $d_{i_2}^o$ be the addresses of the terms in the ${}_oD$-dictionary between which the new term has been inserted. Then the value of i in the address of the new term is $\frac{i_1 + i_2}{2}$. Of course, $i_1 = 0$ if the new term is first on the list and $i_2 = 1$ if it is last. This completes the choice of the address for the new element. Now its synonyms can be inserted in the proper places of the remaining dictionaries. Finally, the translator of the system is updated by incorporating a program that translates the new main term and every synonym into the new address $d_i^o$.

A system should have a provision to delete elements of level zero that are not connected to any other elements of the system. The rules for this procedure are described below in this section. In order to avoid deletion of a newly added element, the introduction of an element of level zero must be accompanied by the introduction of an element of higher level that contains the new zero level term in its verbal statement.

Elements of level higher than zero are constructed as follows. First, the designer of the information system constructs all seven components of the new element. The third and the fourth components are sets of addresses of those concepts that appear in the fifth component $[a_j^m]$ of the new element. All these concepts must be in the system already. The designer chooses the first component a of the address of the new element according to its type. The value of a and all seven components are read into the system. The system determines automatically the maximum $k_1$ of all the second components in addresses contained in the third and fourth component of the new element and it assigns $k = k_1 + 1$ as the second component of the new address. The third component is obtained automatically with the aid of the ${}_kA$ dictionary, as described above for elements of level zero. After this all the dictionaries are updated, and a translator of the new main term and its synonyms into the new address is added. Incorporation of the new element into the system is completed by automatically connecting the sets ${}^kA_i$ and $A_i^k$ with the elements of the system whose addresses appear in these sets.

Elimination of elements from the system is performed as follows. We define a sequence of instances $t_1, t_2, \ldots$ either by specifying a time increment $\Delta t$ between the successive instances or by choosing the number of requests for information which are to be processed between successive values of $t_k$. At every instance $t_k$ one is subtracted automatically from the rescission index of every element in the information. Whenever an answer to a request for information includes the verbal statement $[a_i^k]$ or the reference $\{a_i^k\}$ of an element, one is added to $\overline{a_i^k}$, the rescission index of $a_i^k$. When the users collectively have no interest in an element of the information or when an element becomes commonly known that it will never or seldom be retrieved, eventually its rescission index will become zero. Then an element of the type p, c, s, or t is automatically removed from the system by deleting its components, by removing the main term and synonyms from the dictionaries, and by destroying its translator and all the connections

of this element with other elements of the information. After deletion of an element of the type p, c, s, or t the system examines all the elements of ${}_{o}D$ that were connected to the deleted element. If an element from this collection has no connections to any other elements of the system, it is also deleted.

When the rescission index of an element of type $d_i^k$ becomes zero, the element is converted into an element $d_j^o$. The first two components of $d_i^k$ become components of the new $d_j^o$. This new element is incorporated into the system by the rules described above for incorporating new elements into ${}_{o}D$. At the same time the system constructs all the negative chains of elements that begin with the old element $d_i^k$ and replaces the address $d_i^k$ in the third and second components of all the elements in these chains by $d_j^o$. After this the second components in the addresses of each element in these chains are computed in succession, starting with the elements of lowest level. This computation is performed by the rules that are available for incorporation of new elements of level higher than zero. If the second component in the address of an element $a_j^m$ belonging to one of these chains is not changed, then all the addresses in the elements of the chains that are in negative relation to $a_j^m$ remain the same, provided there is no other negative chain that leads from the element $d_i^k$ to the element $a_j^m$. If, however, the second component of the address $a_j^m$ is changed, then the addresses of the elements in the negative chains that begin with $a_j^m$ are adjusted by the same rules.

A question of any type considered in the previous section can readily be answered by this system. Suppose we have a question of type ($\iota$) about a subject described by the sentence q. The user specifies the class of the question and feeds the actual question into the system. As successive words in the question are read-in, the translator ignores all those words that do not belong to the main dictionary and translates the others into their addresses $a_i^k$, thus generating the set E. The number of elements in E is counted at the same time. Let this number be n. Now all the elements with addresses in E are activated, and they in

turn send activation signals to all the elements in I that are connected to these elements. At the same time a printer is set to accept signals from all the elements in I that have exactly n active connectors in ${}^{k}A_i \cup A_i^k$. The user may specify which components of these elements should be printed or displayed. Similarly a simple procedure can be described for retrieving an answer to any other class of questions. These procedures are analogs of the rules described above for a card file. We omit this reiteration.

The connection between elements need not be physical. It is enough to be able to activate simultaneously all "replicas" of a given element, i.e. to make a concept in the system one indeed. Such an automatic connection can be visualized by assuming that the same address (which is just a code name for an element) whenever it is in the system responds to, say, a specified electromagnetic frequency. Then the required "connections" to a new element would be automatically established once its sensitivity to the respective frequencies is introduced. In this abstract picture the information system would be just a collection of structured elements stored in any order in a certain location with all the necessary relations and connections provided by their sensitivity to appropriate frequencies and their ability to transmit the signals.

## VI. CONCLUDING REMARKS

The preceding discussion is not meant to be a solution to the problem of information retrieval. It is rather just a formulation of the problem with an analysis of a possible approach to its solution. A further study of the problem is needed. It is very likely that a careful examination of the needs of information users will lead to a more comprehensive list of classes of questions than the one discussed in Section III. Additional classes and a need for more precise retrieval may require redefinition of the elements of information. It may be that information should consist of elements with more than seven components and of much more complex structure than that described above. However,

we feel that any approach with some hope of success must concentrate on information instead of documents. This is the main thesis of our report.

Of course, construction of an information system of this type would be rather expensive. Nevertheless, in view of the explosive proliferation of research publications, a great number of which contain only noise, construction of an information system may be less expensive than mere sorting and labeling of documents. If we try to estimate the cost of time spent by professionals while searching through irrelevant documents for desired information; the cost of lost information, i.e. information that is not retrieved when needed; the cost of publishing all the jumble that is presented as original research; and above all, the cost of producing all those mountains of gibberish, then, maybe, we will find that the expenses in professional effort and in money for building an adequate information system are relatively modest. Readily accessible information that includes only significant and original results, we believe, would discourage production of less than mediocre publications. A system that is able to incorporate new results at a sufficiently early date may even lead to reduction of the number of research journals, saving thereby the professional effort of reviewers and editors and also the cost of publishing. In any case, we believe that further study of feasibility and of methods for constructing an information system should be conducted. Examples of information sets should be developed, and the main problems in their construction should be identified for further analysis.

## REFERENCES


[ 1] P. B. Baxendale, "Machine-Made Index for Technical Literature - an Experiment," IBM J. Res. and Dev., 1958, Volume 2, pp. 354-361.

[ 2] H. Borko and M. Bernick, "Automatic Document Classification," J. Assoc. Comp. Mach., 1963, Volume 10, pp. 151-162.

[ 3] H. P. Luhn, "A Statistical Approach to Mechanized Encoding and Searching of Literary Information," IBM J. Res. and Dev., 1957, Volume 1, pp. 309-317.

[ 4] C. K. Van Meter, "A Proposed Chemical Information and Data System," Edgewood Arsenal, 1965, AD 477 110L

[ 5] W. R. Plugge and M. N. Perry, "American Airlines' "SABRE" Electronic Reservations System," Proc. Western Joint Comp. Conf., May 1961, pp. 593-602.

[ 6] J. B. Dennis, "A Position Paper on Computing and Communications," Commun. of the ACM, Volume 11, No. 5, pp. 370-377, 1968.

[ 7] J. O. Harrison, Jr., "A Matrix Technique for Describing Reporting Dependence in Information Systems," Research Analysis Corporation, 1968.

[ 8] T. H. Naylor, J. L. Balintfy, D. S. Burdick, and Kong Chu, "Computer Simulation Technique," John Wiley and Sons, 1967.

[ 9] R. H. Bruck, "A Survey of Binary Systems," Springer Verlag, 1958.

[10] G. L. Peakes, A. Kent, and J. W. Perry, "Progress Report in Chemical Literature Retrieval," Interscience Publishers, Inc. New York, 1957.

Unclassified

Security Classification

**DOCUMENT CONTROL DATA - R & D**

*(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)*

| | |
|---|---|
| 1. ORIGINATING ACTIVITY (Corporate author)<br>U.S. Army Aberdeen Research and Development Center<br>Ballistic Research Laboratories<br>Aberdeen Proving Ground, Maryland | 2a. REPORT SECURITY CLASSIFICATION<br>Unclassified<br>2b. GROUP |

3. REPORT TITLE

INFORMATION AND ITS RETRIEVAL

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

5. AUTHOR(S) (First name, middle initial, last name)

Česlovas Masaitis

| 6. REPORT DATE | 7a. TOTAL NO. OF PAGES | 7b. NO. OF REFS |
|---|---|---|
| Novmeber 1970 | 57 | 10 |

| | |
|---|---|
| 8a. CONTRACT OR GRANT NO.<br>b. PROJECT NO. RDT&E 1T061102A14B<br>c.<br>d. | 9a. ORIGINATOR'S REPORT NUMBER(S)<br>BRL Report No. 1513<br>9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) |

10. DISTRIBUTION STATEMENT

This document has been approved for public release and sale; its distribution is unlimited.

| 11. SUPPLEMENTARY NOTES | 12. SPONSORING MILITARY ACTIVITY |
|---|---|
| | U.S. Army Materiel Command<br>Washington, D.C. |

13. ABSTRACT

Information is here defined abstractly as a set of highly structured elements. This structure induces a two-way classification of the elements and also other relations among them. The definition includes rules for expanding this information by incorporating new elements and for discarding elements that have become obsolete or are no longer needed. The definition of information is geared to its purpose, viz. to provide easy retrieval of known facts that are of interest to a specialist in the field.

A pilot example of such a retrieval system is included, and principles for the physical realization of the system are presented.

It is stated that a clear distinction must be made between a collection of documents and the information they contain. Likewise, the difference between recovering relevant documents and retrieving desired information is emphasized.

DD FORM 1 NOV 68 1473 REPLACES DD FORM 1473, 1 JAN 64, WHICH IS OBSOLETE FOR ARMY USE.

Unclassified

Security Classification

Unclassified

Security Classification

| 14. KEY WORDS | LINK A | | LINK B | | LINK C | |
|---|---|---|---|---|---|---|
| | ROLE | WT | ROLE | WT | ROLE | WT |
| Information<br>Information retrieval,<br>Structure of information,<br>Mathematical model of information | | | | | | |

Unclassified

Security Classification